**GENDER IDENTIFICATION USING SPEECH SIGNAL**

*A project report submitted in partial fulfillment of the requirements for the degree*

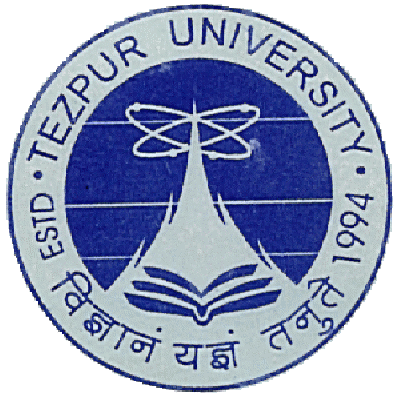
*of*

**MASTER OF COMPUTER APPLICATION (MCA)**

OF

**TEZPUR UNIVERSITY**

2019



*Submitted by*

**HIRAK JYOTI NATH (CSM17033)**

**KM PRAUTI (CSM17013)**

*Guided by*

|  |
| --- |
| **DR. SANGHAMITRA NATH**  PROFESSOR  TEZPUR UNIVERSITY |
|  |

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

**TEZPUR UNIVERSITY**

TEZPUR 784028

ASSAM.

|  |  |
| --- | --- |
| 1tu_logo | DEPARTMENT OF  **COMPUTER SCIENCE & ENGINEERING**  **TEZPUR UNIVERSITY**  TEZPUR 784028  ASSAM. |

CERTIFICATE

This is to certify that the project report entitled **“GENDER IDENTIFICATION USING SPEECH SIGNAL”** submitted as MCA 5th Semester Minor project, was carried out by **Hirak Jyoti Nath(CSM17033) and KM Prauti(CSM17013)** under my supervision. The work has not been submitted to any other institute for the award of any other degree or diploma.

Date:Signature of the supervisor

Place: Tezpur **Dr. Sanghamitra Nath**

Professor,

Department of CSE

Tezpur Univeristy

**DEPARTMENT**

**OF**

**COMPUTER SCIENCE & ENGINEERING**

**SCHOOL OF ENGINEERING**

**TEZPUR UNIVERSITY**

**NAPAAM, SONITPUR – 784028**

**ASSAM, INDIA**

****

**CERTIFICATE**

This is to certify that the project entitled **“GENDER IDENTIFICATION USING SPEECH SIGNAL”** is submitted by **Hirak Jyoti Nath(CSM17033) and KM Prauti(CSM17013)** in partial fulfillment of the requirements for the completion of the degree of **Master of Computer Applications** **5th Semester** is a record of bonafide work carried out by him.

**Dr. Bhogeswar Borah**

**Head of Department**

**Professor**

Date: Department of CSE

Place: Tezpur Tezpur University

**CONTENTS**

# Acknowledgement

1. Abstract
2. Introduction
3. Literature Review
4. Deep Convolutional Neural Network
5. Dataset and Software Libraries
6. Conclusion
7. Future Scope

1. **ACKNOWLEDGEMENT**

The minor project is entitled as “**GENDER IDENTIFICATION USING SPEECH SIGNAL** ” we share a thought of giving ovation to all the people for their support, valuable advice and guidance that has helped us in carrying out this project without their support and guidance it was very hard to finish the project.

First of all, we would like to convey our thanks to Dr. Bhogeswar Borah, Professor and Head of the Department, Department of Computer Science and Engineering, Tezpur University for allowing us to carry out the project “GENDER IDENTIFICATION USING SPEECH SIGNAL”.

At very outset, we take the opportunity to offer our heart-full thanks and sincere gratitude to Dr. Sanghamitra Nath, Professor, Department of Computer Science and Engineering, Tezpur University for offering graceful guidance and support in every step of the project work without which it would have been impossible to fulfill the target.

We would like to express our heart filled thanks to all faculty and technical stafffor their help and support during this project. We would also like to thanks all of our friends for their valuable help, views and constant encouragement throughout the entire project periods.

HIRAK JYOTI NATH (CSM17033)

KM PRAUTI (CSM17013)

MCA 5th Semester

Department of CS

**2. ABSTRACT**

Gender is a critical statistic characteristic of individuals. In this work we attempt to identify human gender using speech signal characteristics such as Mel spectrograms and classifiers such as Convolution Neural Networks. A review of approaches exploiting information from human speech is also presented. Here, highlights of selection of speech features, their processing and the classifiers used for this purpose are discussed.

**3. INTRODUCTION**

Gender identification, language identification, emotion recognition etc. of speech through trained classifiers using speech signal properties has a increased attention in recent years. Accurate classification of above said attributes can be used for other applications based on speech signals such as human-computer interaction, surveillance, speech recognition, speaker recognition, speech coding, language dependent searching, biometrics, gender specific advertising etc.

Human can easily identify the gender of persons by listening their speech but it is a tough task for automated Machine. Here, a brief survey of human gender identification using speech signal is presented. Basic differences between male and female speakers are their pitch and frequencies. So, researchers in early stage used pitch and formant frequencies features of speech to distinguish between the male speaker and female speaker. Since, these features are not robust in case of noise and unvoiced sound then different cepstral domain features are also used for the said purpose.

Few researchers used combination of time domain features like Pitch and cepstral domain features like MFCC, different frequencies etc. and fuse these features together to form a feature vector. Now, these feature vectors are used for training and testing different types of classifiers like neural network, support vector machine, random forest etc. Some researchers also tried to calculate minimum duration of speech required for identifying the gender of a speaker with highest accuracy[1].

In particular, deep convolutional neural networks (CNN)[7] are, in principle, very well suited to the problem of sound classiﬁcation: ﬁrst, they are capable of capturing energy modulation patterns across time and frequency when applied to spectrogram-like inputs. Second, by using convolutional kernels (ﬁlters) with a small receptive ﬁeld, the network is able to successfully learn and later identify spectro-temporal patterns that are representative of different sound classes even if part of the sound is masked (in time/frequency) by other sources (noise), which is where traditional audio features such as MelFrequency Cepstral Coefﬁcients (MFCC) fail. Deep neural networks, which have a high model capacity, are particularly dependent on the availability of large quantities of training data in order to learn a non-linear function from input to output that generalizes well and yields high classiﬁcation accuracy on unseen data.

**4. LITERATURE REVIEW**

At the beginning of this work , we have gone through various research papers based on gender identification using speech signal. We have found several classifiers such as GMM, i-vector, CNN etc which have been used to model the dataset. Compared to GMM[8] and i-vector models, we have seen that CNN have been able to show superior performance in this regard.

That is why we have selected CNN model for classification of speech sample. But CNN needs huge amount of data to provide accurate results. This problem can be solved by making the training set larger by using data augmentation.

**5. Deep Convolutional Neural Network**

The deep convolutional neural network (CNN) architecture proposed in this study is comprised of 3 convolutional layers interleaved with 2 pooling operations, followed by 2 fully connected (dense) layers. The input to the network consists of timefrequency patches (TF-patches) taken from the log-scaled melspectrogram representation of the audio signal. Speciﬁcally, we use to extract log-scaled mel-spectrograms with 128 components (bands) covering the audible frequency range (0-22050 Hz), using a window size of 23 ms (1024 samples at 44.1 kHz) and a hop size of the same duration. Since the excerpts in our evaluation dataset (described below) are of varying duration (up to 14 s), we ﬁx the size of the input TF-patch X to3seconds(128frames),i.e. X ∈R128×128.TFpatches are extracted randomly (in time) from the full log-melspectrogram of each audio excerpt during training as described further down.

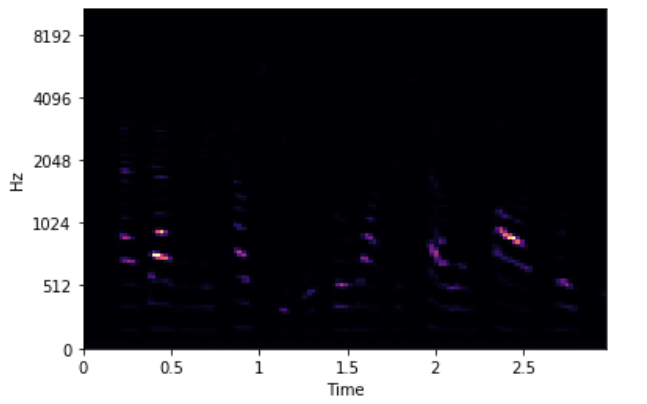


Fig: Melspectrogram of a speech sample

Given our input X, the network is trained to learn the parameters Θ of a composite nonlinear functionF(·|Θ) which maps X to the output (prediction) Z:

Z = F(X|Θ) = fL(···f2(f1(X|θ1)|θ2)|θL), (1)

where each operation f`(·|θ`) is referred to as a layer of the network, with L = 5 layers in our proposed architecture. The ﬁrst three layers, l ∈ {1,2,3}, are convolutional, expressed as:

Z` = f`(X`|θ`) = h(W ∗X` + b), θl = [W,b] (2)

where X` is a 3-dimensional input tensor consisting of N feature maps, W is a collection of M 3-dimensional kernels (also referred to as ﬁlters), ∗ represents a valid convolution, b is a vector bias term, and h(·) is a point-wise activation function. Thus, the shapes of X`, W, and Z` are (N,d0,d1), (M,N,m0,m1) and (M,d0−m0+1,d1−m1+1) respectively. Note that for the ﬁrst layer of our network d0 = d1 = 128, i.e., the dimensions of the input TF-patch. We apply strided max-pooling after the ﬁrst two convolutional layers ` ∈{1,2} using a stride size equal to the pooling dimensions (provided below), which reduces the dimensions of the output feature maps and consequently speeds up training and builds some scale invariance into the network. The ﬁnal two layers, ` ∈ {4,5}, are fully-connected (dense) and consist of a matrix product rather than a convolution:

Z` = f`(X`|θ`) = h(WX` + b), θ` = [W,b] (3)

where X` is ﬂattened to a column vector of length N, W has shape (M,N), b is a vector of length M and h(·) is a point-wise activation function.

The proposed CNN architecture is parameterized as follows:

• l1: 24 ﬁlters with a receptive ﬁeld of (5,5), i.e., W has the shape (24,1,5,5). This is followed by (4,2) strided maxpooling over the last two dimensions (time and frequency respectively) and a rectiﬁed linear unit (ReLU) activation function h(x) = max(x,0).

• l2: 48 ﬁlters with a receptive ﬁeld of (5,5), i.e., W has the shape (48, 24, 5, 5). Like `1, this is followed by (4,2) strided max-pooling and a ReLU activation function.

• l3: 48 ﬁlters with a receptive ﬁeld of (5,5), i.e., W has the shape (48, 48, 5, 5). This is followed by a ReLU activation function (no pooling).

• l4: 64 hidden units, i.e., W has the shape (2400, 64), followed by a ReLU activation function. • l5: 10 output units, i.e., W has the shape (64,10), followed by a softmax activation function.

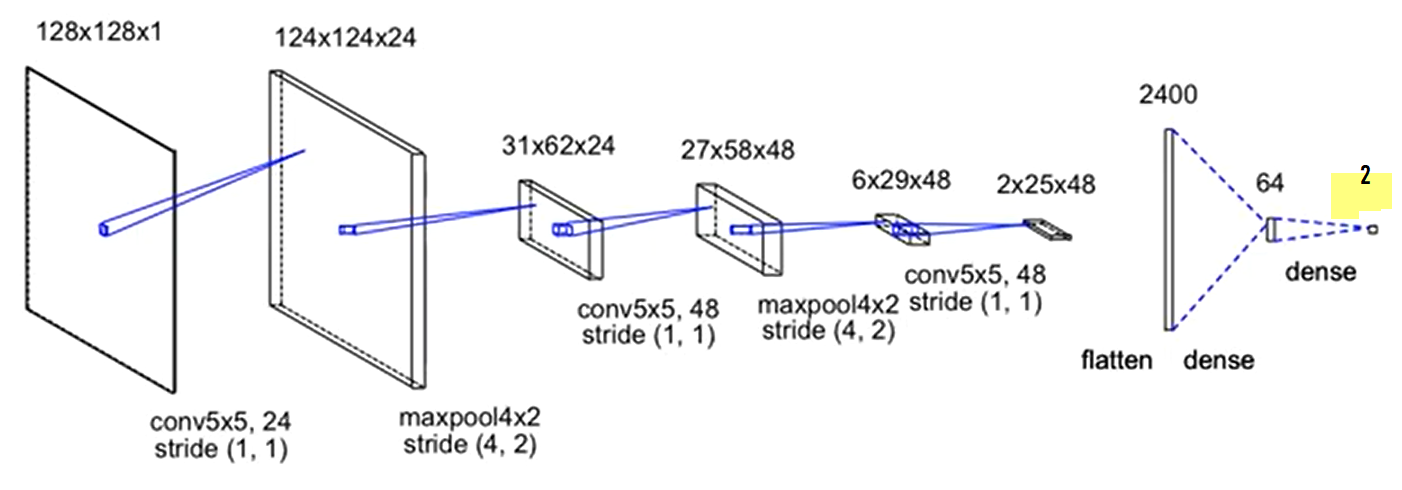


Fig: Convolutional Neural Network Architecture

Our use of a small receptive ﬁeld (5,5) in l1 compared to the input dimensions (128,128) is designed to allow the network to learn small, localized patterns that can be fused at subsequent layers to gather evidence in support of larger “time-frequency signatures” that are indicative of the presence/absence of different sound classes, even when there is spectro-temporal masking by interfering sources.

For training, the model optimizes cross-entropy loss via mini-batch stochastic gradient descent . Each batch consists of 100 TF-patches randomly selected from the training data (without repetition). Each 3 s TF-patch is taken from a random position in time from the full log-mel-spectrogram representation of each training sample. We use a constant learning rate of 0.01. Dropout is applied to the input of the last two layers, l ∈ {4,5}, with probability 0.5. L2regularization is applied to the weights of the last two layers with a penalty factor of 0.001. The model is trained for 50 epochs and is check pointed after each epoch, during which it is trained on random mini batches until 1/8 of all training data is exhausted (where by training data we mean all the TF-patches extracted from every training sample starting at all possible frame indices). A validation set is used to identify the parameter setting (epoch) achieving the highest classiﬁcation accuracy, where prediction is performed by slicing the test sample into overlapping TF-patches (1-frame hop), making a prediction for each TF-patch and ﬁnally

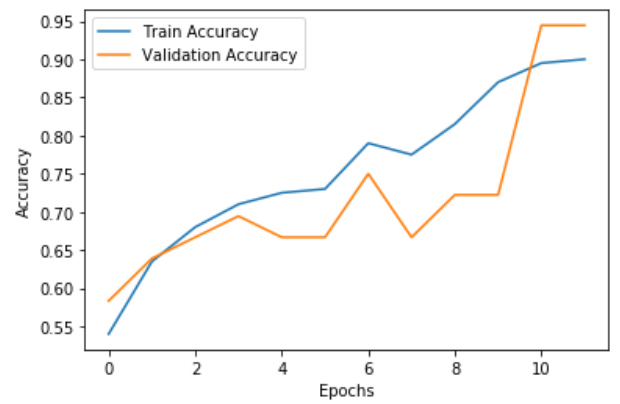


Fig: Showing the accuracy for training and validation datasets

choosing the sample level prediction as the class with the highest mean ouptut activation over all frames. The CNN is implemented in Python.

**6. DATA SET AND SOFTWARE LIBRARIES**

* 1. **DATA SET**

We use the LibriSpeech[4] corpus of audiobook data to train and evaluate models. We use the train-clean-100[4] for training. This data contains audio from controlled environments, no external noise just recording artifacts such as microphone buzz. The LibriSpeech corpus is available free of charge. In this project we have used 88.78 minutes (approx. 1.47hrs) of “clean” speech. Each voice sample format is a .FLAC file. The pre-processed FLAC files have been saved into a CSV file. The CSV file is contained 403 rows and 4 columns.

* 1. **SOFTWARE LIBRARIES**

Python; is an interpreted, interactive, object-oriented, dynamic type, easy to learn and open source programming

Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.

Keras; “is a high-level neural networks library, written in Python and capable of running on top of either TensorFlow or Theano” [9].

TensorFlow™ is an open source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them [14]. TensorFlow's flexible architecture allows you to use GPU or CPU to mainly conducting machine learning and deep neural networks research, but other domains can be adapted easily.

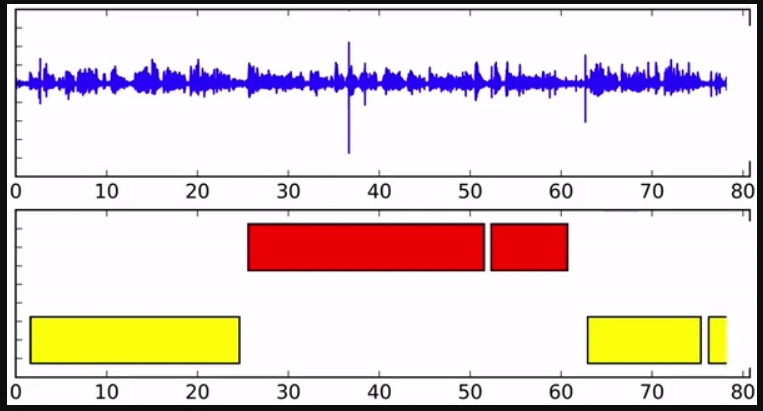
NumPy is the open source fundamental package for scientific computing with Python. It contains powerful capabilities such as N-dimensional array objects, sophisticated (broadcasting) functions, tools for integrating C/C++ and Fortran code, useful linear algebra, Fourier transform, and random number capabilities [10]. By using Numpy arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases. Keras uses Numpy for input data types.

**7. CONCLUSION**

Our work shows that we can use acoustic properties of the voices and speech to detect gender of the speaker. A deep CNN model has been realized to segment speech streams into male and female excerpts and compared to GMM and i-vector models. This positive outcome allows us considering the reliability of this gender detection system sufﬁcient to perform large-scale gender statistic descriptions, offering concrete perspectives for digital humanities.We have shown that the accuracy of the model is 94 % which can be further improved with larger amount of datasets and augmented training set.

**8. FUTURE SCOPE**

In future, we can further enhance this project into a Speaker diarization[5]. Speaker diarization is the process of partitioning an audio stream with multiple people into homogeneous segments associated with each individual, is an important part of speech recognition systems. By solving the problem of “who spoke when”, speaker diarization has applications in many important scenarios, such as understanding medical conversations, video captioning and more.

****

Online speaker diarization on streaming audio input. Different colors in the bottom axis indicate different speakers.

**SOURCE CODE**

import keras

from keras.models import Sequential

from keras.layers import Activation,Dense, Dropout, Flatten, Conv2D,MaxPooling2D

from glob import glob

import numpy as np

import pandas as pd

import random

import matplotlib.pyplot as plt

#Librosa for audio

import librosa as lr

#And the display module for visualization

import librosa.display

data=pd.read\_csv('E:/AI/Gender Recognition of Speaker/training samples/voices/voicesample.csv')

data.head(5)

#Display No. of rows and columns

data.shape

#read data

data\_dir = 'E:/AI/Gender Recognition of Speaker/training samples/voices'

audio\_files = glob(data\_dir + '/\*.flac')

print(len(audio\_files))

# Load the audio as a waveform `y`

# Store the sampling rate as `sr`

y,sr=lr.load(audio\_files[5], duration=2.97)

print(y)

print(sr)

plt.plot(y)

# Let's make and display a mel-scaled power (energy-squared) spectrogram

ps=librosa.feature.melspectrogram(y=y, sr=sr)

print(ps)

ps.shape

# Display the spectrogram on a mel scale

librosa.display.specshow(ps, y\_axis='mel', x\_axis='time')

D=[] #DataSet

for row in data.itertuples():

print(row)

D=[] #DataSet

for row in data.itertuples():

# print(row)

y,sr=lr.load('E:/AI/Gender Recognition of Speaker/training samples/voices/' + row.Filename, duration=2.97)

ps=librosa.feature.melspectrogram(y=y, sr=sr)

if ps.shape !=(128,128):

#print(file)

continue

D.append((ps,row.Class))

print(D)

print("Number of samples:",len(D))

dataset = D

random.shuffle(dataset)

train=dataset[:200]

print(train)

test=dataset[200:]

print(test)

X\_train, Y\_train = zip(\*train)

print(X\_train)

print(Y\_train)

X\_test, Y\_test = zip(\*test)

#Reshape for CNN input

X\_train = np.array([x.reshape((128,128,1)) for x in X\_train])

X\_test = np.array([x.reshape((128,128,1)) for x in X\_test])

print(X\_train)

# One-Hot encoding for classes

Y\_train = np.array(keras.utils.to\_categorical(Y\_train,2))

Y\_test = np.array(keras.utils.to\_categorical(Y\_test,2))

print(Y\_train)

print(Y\_test)

model = Sequential()

input\_shape=(128,128,1)

model.add(Conv2D(24,(5,5), strides=(1,1), input\_shape=input\_shape))

model.add(MaxPooling2D ((4,2), strides= (4,2)))

model.add(Activation('relu'))

model.add(Conv2D (48, (5,5), padding = 'valid'))

model.add(MaxPooling2D ((4,2), strides = (4,2)))

model.add(Activation('relu'))

model.add(Conv2D (48, (5,5), padding = 'valid'))

model.add(Activation('relu'))

model.add(Flatten())

model.add(Dropout( rate = 0.5))

model.add(Dense(64))

model.add(Activation('relu'))

model.add(Dropout(rate= 0.5))

model.add(Dense(2))

model.add(Activation('softmax'))

model.compile(

optimizer="Adam",

loss = "categorical\_crossentropy",

metrics=['accuracy'])

history=model.fit(

x = X\_train,

y = Y\_train,

epochs = 12,

batch\_size = 40,

validation\_data = (X\_test, Y\_test))

score=model.evaluate(

x=X\_test,

y=Y\_test)

print('Test loss:', score[0])

print('Test accuracy:', score[1])

model.summary()

plt.plot(history.history['acc'], label='Train Accuracy')

plt.plot(history.history['val\_acc'], label='Validation Accuracy')

plt.xlabel('Epochs')

plt.ylabel('Accuracy')

plt.legend()

**REFERENCES:**

1. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classiﬁcation Justin Salamon and Juan Pablo Bello
2. <https://www.youtube.com/watch?v=GNza2ncnMfAR>
3. A Survey Paper on Gender Identification System using Speech Signal by

Mohit Kumar Mishra, Arun Kumar Shukla

1. LibriSpeech ASR Corpus is an Open Speech and Language Resource Website (<http://www.openslr.org/12/>), train-clean-100 contains training set of 100 hours of “clean” speech. Building a Speaker Identification System from Scratch with Deep Learning by by Oscar Knagg
2. Accurate Online Speaker Diarization with Supervised Learning by Chong Wang, Research Scientist, Google AI (<https://ai.googleblog.com/2018/11/accurate-online-speaker-diarization.html>)
3. Nptel MOOCS course on Deep Learning (IITKGP) by Prof. Prabir Kumar Biswas
4. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
5. Voice Gender Recognition Using Deep Learning by Mucahit Buyukyilmaz1,\* and Ali Osman Cibikdiken2
6. Keras, Chollet, François, 2015, <https://github.com/fchollet/keras>
7. S. van der Walt, S.C. Colbert, G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science & Engineering, 13, 22-30, 2011, doi:10.1109/MCSE.2011.37